DOCUMENT RESUME

ED 333 007 TM 016 630

AUTHOR Banerji, Madhabi
TITLE Predictive Properties of the Gesell School Readiness Screening Test within Samples from Two Treatment Contexts.
PUB DATE Apr 91
NOTE 27p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 3-7, 1991).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Ability Grouping; Academic Achievement; Cognitive Development; Comparative Testing; High Achievement; *Kindergarten Children; Low Achievement; *Predictive Validity; *Preschool Children; Primary Education; *School Readiness; *Screening Tests; *Student Placement; Test Results; Test Use; Track System (Education)
IDENTIFIERS *Gesell School Readiness Test

ABSTRACT

The predictive properties of the Gesell School Readiness Screening Test (GSRT) were examined, taking into account the stated purposes of the test and the context of test use. Two samples were used: (1) a control sample of 55 students (21 males and 34 females) whose GSRT scores were not used for placement or tracking; and (2) a treatment sample of 70 children (32 males and 38 females) attending schools that were pilot testing the GSRT along with an extra-year tracking component for students with low developmental age scores on the GSRT (DAs). Two subgroups of the control sample were identified: those with low DAs who were not placed in a 2-year kindergarten, and those with high DAs who were placed in a 1-year kindergarten. Three subgroups of the treatment sample were identified: those with low DAs who were placed in a 2-year program, those with low DAs who were not placed in a 2-year program, and those with high DAs who were placed in a 1-year program. Student achievement was determined in grade 1. The GSRT was found to be a modest predictor of subsequent school achievement, and the DA was a better predictor than chronological age. Predictive validity coefficients alone were not sufficient to support the use of the test in placing and extra-year tracking of students with low DAs. In addition, the classification properties of the GSRT were found to be rather weak, with a large proportion of false negative errors in identifying at-risk students on the low end of the scale. The GSRT is more suitable as a screening test than as a student placement test. Seven tables, three figures, and a 34-item list of references are included. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

MADHABI BANERJI

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED333007

Predictive Properties of the Gesell School Readiness Screening Test within Samples from Two Treatment Contexts

by
Madhabi Banerji
Pasco County School District

Paper Presented at the Annual Conference of the American Educational Research Association
Chicago, April 7, 1991

TM016380

BEST COPY AVAILABLE

Predictive Properties of the Gesell School Readiness Screening Test within Samples from Two Treatment Contexts

Introduction

This study was conducted in a mid-size school district (pop. 33,000) in the state of Florida where the Gesell School Readiness Screening Test (GSRT) was in use as a part of a program designed to identify students at risk of school failure. Identification of at-risk students at the kindergarten level was legislated in public school systems in Florida by means of the Florida Primary Education Program (PREP) legislation passed in 1979. Several school systems incorporated the use of the Gesell developmental tests in their kindergarten and first grade programs to meet the above requirement (See 1989 Survey of Alternate Placement Programs in Florida, Division of Public Schools, Florida Dept. of Education, 1989).

The GSRT is marketed accompanied by specific placement recommendations related to test outcomes. Specifically, students who show low performance on the test are considered developmentally delayed for their age (immature) and are recommended by the test developers for two-year kindergarten placement prior to beginning first grade (Gesell Institute of Child Development, 1980). Typically, schools that employ the Gesell tests to identify at-risk students, also follow the two-year tracking recommendation given by the authors.

Existing empirical research to support application of the GSRT in the manner recommended is limited. In light of the increased use of the tests in combination with the two-year treatment recommended for students identified as delayed, several authors in measurement and early childhood education recently pointed to the need for research on the Gesell tests (Naglieri, 1985; Kaufman, 1985; Bradley, 1985; Waters, 1985; Shepard and Smith, 1986; Meisels, 1987; 1988; Bredekamp and Shepard, 1989).

The purpose of this study was to examine the predictive properties of the GSRT taking into consideration the stated purposes of the test (by its developers), the programmatic context of test use, and actions related to the test. The study was conducted on two samples taken from structurally distinct educational programs where the test was employed. In the first, the two-year placement

Acknowledgement: The author thanks Jeffrey D. Kromrey, of the University of South Florida , for his helpful comments on an earlier version of this paper.

recommendation of the Gesell Institute was followed for students identified on the test as immature (the treatment program); in the second, the GSRT was used, but there was no differential programming for students assessed as mature and immature on the GSRT (the control program).

The Gesell School Readiness Screening Test: Content and Underlying Assumptions

The Gesell School Readiness Screening Test (GSRT) represents a screening portion of a larger battery of tasks, entitled the Gesell Preschool Tests (GPT), which is designed to measure school readiness in children in the age range of 2-6 years. The GSRT is an individually administered test composed of the following eight subtests.

Cubes: a block building task using ten one-inch wooden cubes. The child's spontaneous play with blocks is followed by building with blocks following demonstrations by the examiner (e.g. vertical tower, train, bridge).

Copy Forms: a paper and pencil task requiring a child to copy lines and progressively complex geometric shapes (e.g. square, triangle, diamond, cube).

Incomplete Man: a paper and pencil drawing task requiring a child to complete the stick figure of a man by filling in the missing parts.

Writing Name: the child is asked to write his/her name and address.

Writing Numbers: the child is asked to write the numbers 1-20.

Animals: the child is asked to name as many animals as possible in one minute.

Interview: eight personal questions are asked regarding mother, father, siblings, CA, etc.

Interests: the child is asked four open-ended questions about what he/she likes to do best at school and at home, indoors and outdoors.

The GSRT can be administered in 20-30 minutes. Developmental Age (DA) which results from the administration of the GSRT, is assigned in years and determined using norms. In scoring the test, examiners also consider how a child attacks the tasks, facial expressions, and factors such as left-to-right orientation. DAs assigned on the subtests are used to generate a Developmental Profile on each child tested. An overall DA is then estimated by the examiner based on the profile.

DA, as measured by the GSRT, is claimed to be a better predictor of school performance than Chronological Age (CA), the customarily used criterion for school entrance (The Gesell Institute of Child Development, 1980; Haines, Ames and Gillespie, 1980). The authors promote the use of the test in classifying

mature and immature children prior to kindergarten entrance. They recommend that students identified as immature on the test be placed in two-year programs prior to entering first grade. The extra year may be provided to the child in kindergarten, a pre-first grade, or at home. The additional time is expected to enable initially delayed students attain the required level of maturity to successfully cope with academic demands of later grades. The Gesellian view that school readiness is primarily a function of biological development is termed the maturationist view in early childhood literature (Shepard and Smith, 1985; Karweit, 1988). This view does not directly address the role of experiential factors in bringing about development in children. The Gesell Institute, therefore, makes no detailed recommendations on the content of the school curriculum during the two year tracking of initially delayed students.

Literature on early childhood "readiness" testing and "screening" provides different definitions of each of these functions. An operational definition of an early childhood "developmental screening" test has been given by Meisels (1983) using the following domains.

Visual-Motor/Adaptive: Items that examine fine motor control, eye-hand coordination, the ability to remember visual sequences, to draw two-dimensional visual forms, and to reproduce three-dimensional visual structures.

Language and Cognition: Items that focus on language comprehension; verbal expression and articulation, and the ability to reason, count and repeat auditory sequences.

Gross Motor and Body Awareness: Items that examine balance, large motor coordination, and the ability to imitate body positions from visual cues (p. 3).

In describing measures of early childhood development, Miesel's (1983) further states that the intent of a developmental screening test (operationalized as above), is to provide a brief assessment of a child's developmental abilities to determine whether a child should undergo further assessments to diagnose abnormalities. The process of screening is the initial step that must be followed by in-depth diagnostic assessments to identify students for placement in special education programs. Screening tests are required to have developmental content, predictive validity, reliability and norms. Miesels (1987) makes a distinction between developmental screening tests and school readiness tests. According to him, school readiness tests are criterion-referenced measures that should closely reflect the content of the curriculum taught in school. The purpose of readiness testing is to ascertain specific curriculum-related strengths and weaknesses for instructional planning. Neither readiness nor developmental screening tests are designed to be used for diagnosis or placement.

If the previously given definitions are used as a framework to categorize the GSRT, it becomes obvious that the latter possesses the properties of a developmental screening test, rather than a test of school readiness. Tasks of the GSRT that fall into the Visual-Motor domain described are Copy Forms, Incomplete Man, Cubes and the writing items. The Interview, Interests and Animals subtests are oral language tasks that obviously belong in the Language and Cognition domains. Although gross motor tasks are absent in the GSRT, the full battery, as represented by the GPT, does include a complement of motor tasks that would closely match the motor domain described in the definition given by Miesels.

The above discussion makes evident that although the GSRT possesses the content and features of a developmental screening test, its authors go a step beyond the standard applications of a screening test by promoting its use as a diagnostic and placement tool. Because of that very recommendation, however, a more rigorous empirical validity check is needed for the GSRT than is commonly warranted for developmental screening tests.

Research Evidence on the GSRT

Existing predictive validity studies on the Gesell tests have used grade placement (Ames and Ilg, 1964), standardized achievement measures (Kaufman and Kaufman, 1972; Popovics, 1982; May 1986; Graue and Shepard, 1988); teacher judgments (Graue and Shepard, 1988); and referrals to remedial programs (Wood, Powell and Knight, 1984) as criterion measures. Aside from the last study listed, which used a discriminant analysis technique, all others reported zero-order correlations between predictor (DA) and criterion.

Predictive validity coefficients in some of the studies are in the high range -- 0.74 (Ames and Ilg, 1964) and .64 (Kaufman and Kaufman, 1972). Other studies report rather low coefficients-- .28 (Popovics, 1982), .11 (May, 1986) and .20-.40 (Graue and Shepard, 1988). The lack of consensus in the results of these studies may have been caused by several factors. Firstly, the nature and length of the instructional intervention, with or without the extra year factor, during the time that lapsed between the measurement of the predictor and criterion was found to vary from study to study. Secondly, the definition of the Gesell DA was also found to differ in different studies--some studies selectively used one or two tasks, while others used more. Finally, the criterion measures used were qualitatively different in different studies, affecting the comparability of the reported coefficients. The largest coefficient (.74), has been criticized as being spuriously high as it may have been influenced by extreme scores (Bradley, 1985) or other

situational variables (Kaufman and Kaufman, 1972).

Interpretations of the classification accuracy of the Gesell tests have been conducted based on the Wood et al (1984) study. Here, chronological age and DA were compared as predictors for students at risk, where being at risk was measured based on teacher referrals for special education placement. Statistical significance in prediction was demonstrated by the authors for DA over CA. In terms of accuracy of placements, Shepard and Smith (1985) analyzed the same data and reported a high agreement rate for successful students, but a 50% misclassification rate for students labeled at-risk. Shepard and Smith (1985) also noted that in the Wood et al study, testing of DA was done at the same time as the criterion measurement, making it a concurrent validity study rather than a predictive one.

The .64 correlation from the Kaufman study has been interpreted as equal to a standard error of measurement of six months (Graue and Shepard, 1988). As a difference of six months is used to decide whether or not a child will be placed in a two-year program, the authors pointed out that the apparently large coefficient may not necessarily ensure accurate placement.

Existing research on the Gesell tests, in particular on the GSRT as it was being locally used, was found to be not only limited but rather fragmentary. In addition, the most favorable predictive evidence was found to be criticized for methodological limitations. The present study sought evidence to test hypotheses on the predictive validity of the GSRT in comparison with chronological age, taking into account relevant treatment-related variables. The usefulness of the GSRT in making classification decisions was also examined.

Methodological Rationale

According to the *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 1985), test validation is inseparable from the social/practical context of test use. In the absence of existing evidence of a test's validity, as was true in the case of the GSRT, the *Standards* (1985) state that the user must assume responsibility in justifying inferences made following the use of the test. Messick (1989) defines test validity as a judgement of the "degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores" (p. 13, emphasis in original).

The same approach was adopted in the design of the present study. Because the GSRT is not accompanied by adequate validity evidence from the test developers, the study was guided by the user rationale. Empirical evidence was gathered from three different perspectives that were rooted in the theoretical assumptions underlying the test. Each approach to test validation is defined here as it was applied in the study. Additionally, a rationale is presented for each to justify its inclusion in the overall validation plan.

Incremental Validity: The first task was to test basic assumption made by the authors that DA correlates positively with standard measures of school achievement, and if used as a criterion for school entrance in addition to chronological age, it would improve the prediction of achievement that could be made on the basis of chronological age alone. *Incremental Validity* of a test refers to the increment in prediction of the criterion obtained by using the test as an added predictor in the prediction equation. Incremental validity is reflected in the magnitude of multiple correlation coefficients before and after the predictor of interest is added to the equation. Sechrist (1963) recommended that incremental validity evidence be sought for tests intended for applied, predictive use. Both the criterion-related validity, as indicated by zero-order correlations of DA with achievement, as well as incremental validity of DA over CA, were investigated independently in samples of students from the treatment and control programs in this study.

Treatment Validity: Use of the GSRT is accompanied by a treatment recommendation having implications for the duration of time between the measurement of predictor and criterion, although not specifically for the instructional content of the intervening program. *Treatment validity* is defined here as the observed efficacy of treatment tied to the test. The importance of assessing situational and subgroup characteristics that might influence predictor-criterion relationships has been emphasized by several authorities (Cronbach and Meehl, 1955; Guion, 1976; Messick, 1989). Although an important focus of this investigation was to test predictive hypotheses regarding the GSRT, effects of variables related to educational intervention given to students classified on the test as mature, immature and borderline, needed to be considered in the interpretation of validity outcomes. The validity of the complete test-treatment package was therefore examined by comparing achievement of subgroups of students tested on the GSRT but exposed to different treatments.

Predictive Utility: *Utility* is a term that is used to connote the usefulness of a test intended for placement in making accurate classification decisions (Messick, 1989). This property is related to the test-criterion correlation (predictive validity)

but also depends on factors such as the theoretical proportion in the student population that would be selected in the absence of the test (base rate), and the proportion selected by the test for placement (the selection ratio). The utility or classification accuracy of the GSRT was examined. The proportion of successful classifications over and above the base rate was computed using Taylor-Russell tables (Taylor and Russell, 1939).

Validation Questions

Following the above rationale, the questions for the study, investigated independently in treatment and control samples, were the following.

1. Does Developmental Age (DA) as measured on the GSRT, correlate positively and significantly with achievement measured in first grade ?
2. Does DA substantially and significantly add to the prediction of achievement in first grade, when used in conjunction with Chronological Age (CA)?
3. Is the performance of students assessed as mature and immature on the GSRT significantly different when the treatment recommended by the test developers is provided as compared to when it is withheld?
4. Using the obtained zero-order correlations of DA and achievement, what is the predictive utility of the test as interpreted by the Taylor-Russell tables?

Method

Subject Characteristics

All children entering the school system in kindergarten are required to be chronologically 5 years old on September 1.

The control sample was drawn from 3 schools of the 1984-85 academic year. In each of these schools, students entering one kindergarten class were tested on the GSRT as a part of a training program that was under way to prepare the school staff for implementing the test on a regular basis in subsequent years. Thus, while these students were tested on the GSRT, their scores were not used for kindergarten placement or tracking. All students were placed in the same level of kindergarten, irrespective of measured differences in DA, and they moved on to first grade the next year. Although each child's DA was initially assessed by examiner trainees in the control program, each assessment was monitored closely and corrected by the consultant trainer as needed. Students who were retained in kindergarten or first grade were deleted from the sample as these students would have contaminated the "extra year" factor that was being controlled here. The control sample (n=55) was nonrandom. Characteristics of the subjects with respect to CA and DA at the time of school entrance, gender, and socioeconomic status (SES) are shown in Table 1. SES was measured based on whether a child was in the subsidized school lunch program. Ethnically, all

but three students in the control sample, were white. Two subgroups within the control sample were identified for examination of treatment validity. These were students with (1) Low DAs but not placed in two-year kindergarten; and (2) High DAs and placed in one-year kindergarten. Mean DAs and CAs of the subgroups are also shown in Table 1.

Table 1. Subject Characteristics - Control Sample

Attribute	Mean	SD	Range
Total Group:			
1. Developmental Age (Points)	9.12	1.4	6-12
Equivalent in years	5-5 1/2 years		4 1/2-6 years
2. Chronological Age in Months	67.7	4.79	63-73
3. Sex	**Frequency**	**Percent (%)**	
Females	34	59.6	
Males	21	36.8	
4. Race			
White	52	91.2	
Black	1	1.8	
Hispanic	1	1.8	
Asian/Pacific Islander	1	1.8	
5. Socioeconomic Status			
On free/reduced lunch program	10	18.2	
On paid lunch program	45	81.8	

Subgroups:	Low DA but not placed			High DA (above 5 years)		
	Mean	SD	n	Mean	SD	n
Developmental Age	7.35	.78	17	9.92	.71	38
Chronological Age	66.88	4.04	17	68.08	5.09	38

Total Group n = 55

The treatment sample was drawn from the 1985-86 school year from four schools that were piloting the GSRT testing along with the extra-year tracking component for students with low DAs. There were three levels of kindergarten in the treatment program, one with the two-year program for delayed students (K1); the others were one-year programs for students assessed as ready (K2, K3). The curriculum in K1 focused on social and motor development with increasing emphasis on academic readiness in the K2 /K3 levels. The K2 and K3 level curricula were not markedly different in content or activities and the same report card was used in both the upper levels to record student progress. K3 students were those who initially started with the highest DAs and were given enrichment activities if the teacher judged it necessary. The sample from the treatment program (n=70) was drawn randomly from each level (K1, K2, K3) of the pilot schools. Composition of the treatment sample with respect to CA, DA, SES, race

and the number in the upper and lower levels of the treatment program is shown in Table 2 . In this sample, three subgroups of students were identified for examination of treatment effects: those with (1) Low DAs and placed in two-year program (2) Low DAs , not placed in two-year program (3) High DAs with one -year program. Analyses were conducted accordingly.

Table 2. Subject Characteristics - Treatment Sample

	Attribute	Mean	SD	Range
	Total Group:			
1.	Developmental Age (Points)	8.18	1.89	3-12
	Equivalent in years	5 years		3 1/2-4 to 6 years
2.	Chronological Age in Months	65.7	3.78	57-73
		Frequency	**Percent (%)**	
3.	Sex			
	Females	38	54.3	
	Males	32	45.7	
4.	Race			
	White	67	97.1	
	Black	1	1.4	
	Hispanic	1	1.4	
	Asian/Pacific Islander	1	1.8	
5.	Socioeconomic Status			
	On free/reduced lunch program	18	25.7	
	On paid lunch program	52	74.3	

Subgroups:	Low DA and placed			Low DA but not placed			High DA (above 5 years)		
	Mean	SD	n	Mean	SD	n	Mean	SD	n
Developmental Age	6.09	1.44	21	7.75	.45	16	9.84	.78	33
Chronological Age	64.70	3.45	21	65.80	3.81	16	66.32	4.07	33

Total Group n = 70

As is evident, not all students in the Low DA category were placed in two-year kindergartens. Some were moved to the higher levels based on teacher judgements of their readiness. Others were placed in K2 or K3 due to parental request.

For both treatment and control samples, the final data consisted of those students on whom complete achievement data were available in first grade.

Figure 1 describes diagrammatically the path followed by students in the treatment and the control programs. Given the Gesell philosophy, the critical treatment variable to consider in the two programs was the different amounts of time given to students assessed as mature and immature before first grade. It was recognized that other curriculum factors would also naturally have an impact on later performance of the students. These differences were not possible to control in the field. Documentary evidence on staff development programs given to teachers in the treatment and control schools, the kindergarten curriculum guides used, the recommended classroom strategies, and the report cards used in the two programs indicated that the curriculum content of the two programs was not markedly different.

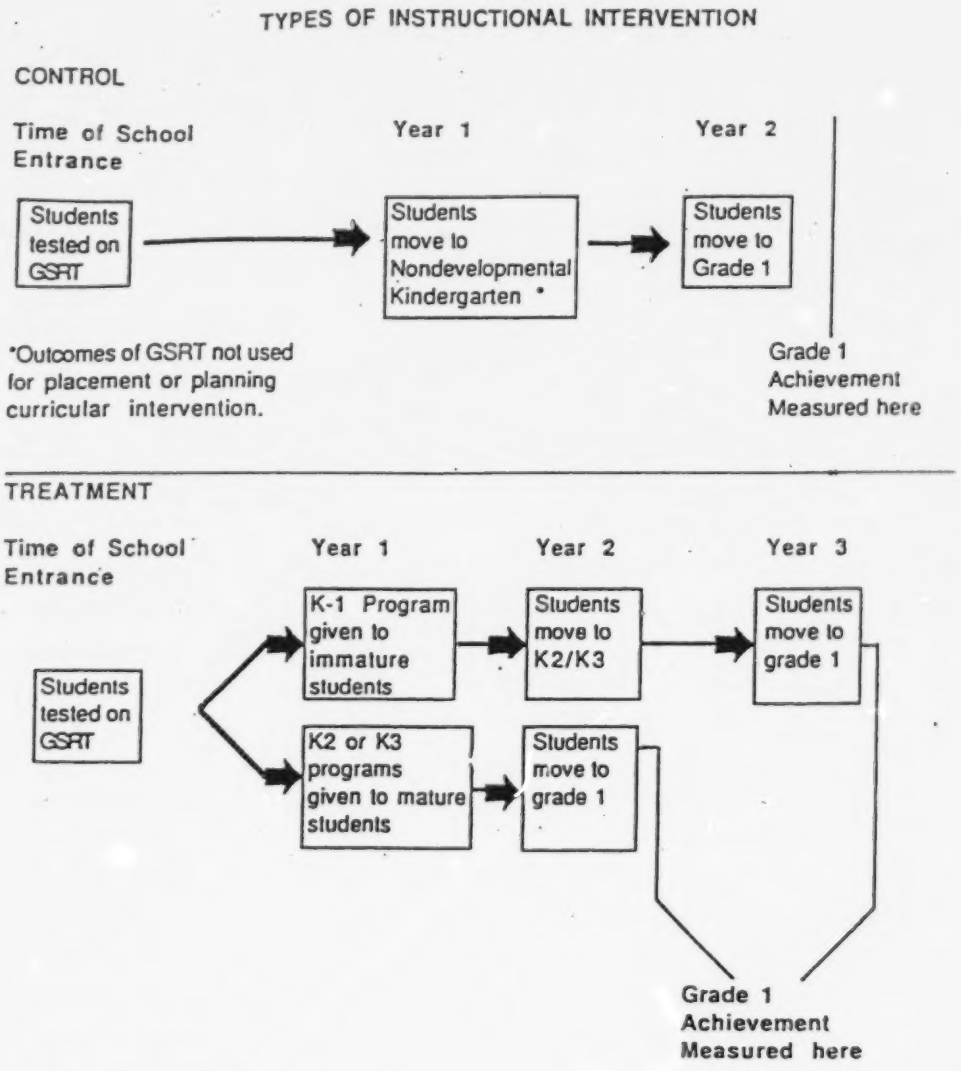

TYPES OF INSTRUCTIONAL INTERVENTION
CONTROL
Time of School Entrance
Year 1
Year 2
Students tested on GSRT
Students move to Nondevelopmental Kindergarten *
Students move to Grade 1
*Outcomes of GSRT not used for placement or planning curricular intervention.
Grade 1 Achievement Measured here
TREATMENT
Time of School Entrance
Year 1
Year 2
Year 3
Students tested on GSRT
K-1 Program given to immature students
Students move to K2/K3
Students move to grade 1
K2 or K3 programs given to mature students
Students move to grade 1
Grade 1 Achievement Measured here

Figure 1. Treatment and Control Program Structures

For the purpose of the study, classification of students into mature and immature categories (Low and High DAs) was conducted using the cutoff score on the DA scale that guided placement decisions in the district. This was a DA of 5-5 1/2 years and greater. This cutoff was in approximate agreement with that recommended by the test authors.

A comparison of the data in Tables 1 and 2 indicates that the treatment and control samples were somewhat different in composition. Equivalence between groups was deliberately not sought for two reasons. First, attempts to match would have resulted in undesirably small sample sizes. Secondly, the intent of the research was to study predictive properties of the GSRT within samples exposed to different treatments rather than to demonstrate between-group differences that result from the different programs. There was adequate variability in DA and achievement within each sample to allow for an examination of predictive validity.

Predictors

DA was operationalized as the overall score on the eight subtests of the GSRT measured in the beginning of the kindergarten year. DAs are assigned in years and can take on intermediate values (e.g. 4 -4 1/2 years). To enable statistical treatment of data, a point system was used to convert the DAs as shown below. This system was adapted from Kaufman (1971) and enabled all possible values of DA to be used for analysis.

Assigned D.A.	3	3-3 1/2	3 1/2	3 1/2-4	4	4-4 1/2	4 1/2	4 1/2-5	5	5-5 1/2	5 1/2	5 1/2-6	6	6-7
Points	0	1	2	3	4	5	6	7	8	9	10	11	12	13

SCALE CONVERSION TABLE

In the school district where the study was performed, the test was administered by trained examiners. In a study of examiner accuracy using expert examiners from the Gesell Institute as the standard, fairly high agreement was found on overall DA's assigned by pairs of examiners examining the same child (Kendall's tau=.85; Pearson r=.91, p<.001). Disagreements were found to be more frequent at the subtest level rather than on the overall score. As only the latter was used for making placement decisions, the subtest disagreements were not considered to be of consequence. Internal consistency reliability estimate for the test was found to be high (.89) with inter-subtest

correlations ranging from .35 - .73. Test-retest reliability, again using the overall DA given on two occasions by the same examiner to a child with a two-week time delay, was found to be high-- Kendall's tau=.82, Pearson r= .90, p< .001 (Banerji, 1990).

Chronological Age (CA), the second predictor in the study, was measured in months at the time of school entrance.

Criterion Variables

School success was measured as performance at the end of first grade on the Total Reading and Total Mathematics batteries of the Stanford Achievement Test (7th Ed.), Primary level 1. Scaled scores were used for each variable. A combined measure of first grade performance was computed by averaging Normal Curve Equivalents (NCE) in reading and math. (NCEs are standard scores with a mean of 50 and a standard deviation of 21).

Analysis

All analyses were conducted using SPSS PC+ software (version 2.0). Parallel analyses were done in the treatment and control samples. Initially, Pearson correlations were run with DA, chronological age (CA) and first grade achievement, to obtain criterion-related validity coefficients. Secondly, incremental validity of DA over CA was examined using hierarchical multiple regressions, where DA was entered after CA, the initial school entrance criterion. The outcome of the hierarchical regression was evaluated using stepwise procedures to eliminate the predictor that reduced statistical significance of results. To examine treatment validity, achievement means were compared using ANOVAs and Scheffe comparisons among relevant subgroups in the treatment and control samples.

Lastly, the classificational properties of the GSRT were examined based on bivariate scatterplots of DA and overall first grade achievement. This analysis was conducted on separate distributions of reading and math scores with little difference in the results. For this reason, only the outcomes with overall first grade achievement are reported here.

The base rate on the standardized achievement test, or the proportion of students assessed on the criterion as at-risk in the absence the GSRT, was identified as 30 % -- which corresponded with the proportion of students performing below the 30th percentile (NCE=40) on the criterion (Stanford Achievement Test). The same cutoff is used in the district for the identification of at-risk students for special referrals and programming. Selection ratios were

computed using the empirically obtained scatterplots. Then, the Taylor-Russell tables were consulted to determine the percent of successful classifications that can be expected with the obtained correlations between DA and achievement, the calculated selection ratio, and the district's base rate. Percent of correct classifications in the empirical distributions were compared to the percents of errors (false positives and false negatives).

Results

Criterion-related Validity

Pearson correlations of DA with CA at the time of school entrance and first grade achievement are presented in Table 3. In both samples, DA was found to correlate moderately with CA, however the correlations were not significant at the .05 level. The range of obtained correlations of DA with achievement were also positive but modest. In the control sample, coefficients were smaller than those in the treatment sample, .24- .29 as compared to .30 -.39. This difference may have had to do with the greater range of DAs in the treatment sample. Further, the correlations in the treatment sample were significantly nonzero while those in the control sample were not, an outcome that may have been caused by the smaller sample size in the latter group.

Table 3. Correlation Matrix of DA, CA and Achievement Variables

Treatment Sample:	DA	CA	Reading	Math	First Grade Achievement
DA	1.0000				
CA	0.2768	1.0000			
Reading	0.3056**	0.0927	1.0000		
Math	0.3934**	0.2077	0.6276**	1.0000	
First Grade Achievement	0.3775**	0.1700	0.9173**	0.8694**	1.0000

n = 70

Control Sample:	DA	CA	Reading	Math	First Grade Achievement
DA	1.0000				
CA	0.3038	1.0000			
Reading	0.2401	0.1160	1.00		
Math	0.2801	0.1904	0.7264**	1.00	
First Grade Achievement	0.2933	0.1487	0.8951**	0.9468**	1.00

** Significant at .001 level

Incremental Validity

Tables 4 and 5 present the hierarchical regressions conducted to examine incremental validity of DA over CA in the treatment and control samples, respectively. Separate analyses were conducted using Reading, Math and overall first grade achievement as the dependent variables.

In the sample from the control program, CA entered first yielded rather small R squared values that were statistically insignificant in all the analyses (R squared values ranged from .014 -.016). Entered second, DA significantly improved the R squared values as well as the statistical significance of the outcomes, with values ranging from .26 -.32. Increments in predictable variance ranged from .06 to .10. In all the analyses, CA was eliminated from the equations following the stepwise procedure.

Table 4. Incremental Validity: Hierarchical Multiple Regression of First Grade Achievement on CA and DA in the Control Sample Followed by Stepwise Analysis

	Entering Variable	Beta	R Square	Rsq. Change	F	P
DV Math						
Hierarchical	1. Age	0.118	0.014		0.745	0.392
	2. DA	0.264	0.082	0.068	2.320	0.108
Stepwise Out	3. Age		0.058		3.239	0.078
		R = .239		Adjusted R Squared = .039		
DV Reading						
Hierarchical	1. Age	0.127	0.016		0.863	0.357
	2. DA	0.306	0.108	0.092	3.139	0.052
Stepwise Out	3. Age		0.078		4.512	0.038
		R = .280		Adjusted R Squared = .061		
DV Grade 1 Achievement						
Hierarchical	1. Age	0.118	0.014		0.76	0.387
	2. DA	0.319	0.113	0.099	3.32	0.044
Stepwise Out	3. Age		0.086		4.98	0.030
		R = .293		Adjusted R Squared = .069		

DV = Dependent Variable
n = 55

Table 5. Incremental Validity: Hierarchical Multiple Regression of First Grade Achievement on CA and DA in the Treatment Sample Followed by Stepwise Analysis

	Entering Variable	Beta	R Square	Rsq. Change	F	P
DV Math						
Hierarchical	1. Age	0.20	0.043		3.06	0.085
	2. DA	0.36	0.165	0.122	6.63	0.002
Stepwise Out	3. Age		0.155		12.45	0.001
		R = .393		Adjusted R Squared = .142		
DV Reading						
Hierarchical	1. Age	0.09	0.009		0.59	0.445
	2. DA	0.30	0.093	0.084	3.45	0.037
Stepwise Out	3. Age		0.090		7.00	0.010
		R = .305		Adjusted R Squared = .08		
DV Grade 1 Achievement						
Hierarchical	1. Age	0.17	0.028		2.02	0.159
	2. DA	0.36	0.147	0.119	5.78	0.005
Stepwise Out	3. Age		0.143		11.30	0.001
		R = .377		Adjusted R Squared = .129		

DV = Dependent Variable
n = 70

In the treatment sample, the results were rather similar. Entering CA first yielded R squared values of .009 in Reading and .04 in Math, both of which were statistically insignificant. Entered second, DA added relatively large increments to the predictable variance (R squared values increased by .08 to .12). DA was also the statistically significant predictor. In all the analyses, CA was dropped from the equations following stepwise analyses.

As is evident from the adjusted R squared values reported in the tables, percent of variance predicted with only DA left in the equations ranged from 4 - 7 % in the control sample and from 8- 14 % in the treatment sample. Relative to CA, DA was the better predictor, irrespective of the type of intervention program given to students. However, overlap in variance between DA and achievement was rather modest.

Treatment Validity

Achievement of students showing initially different DAs (High vs. Low) but who were not given differential treatment is compared in Table 6. These are the subgroups within the control sample. Performance in Math and Reading in first grade was not very disparate for these groups although the delayed students had not received the recommended extra year in kindergarten. A 10 point difference in NCE was observed on the combined first grade measure, but this difference was not significant at the .05 level.

Table 6. Treatment Validity: ANOVAs showing Subgroup Performance in First Grade - Control Sample

	Sum of Squares	df	Mean Square	F	p
ANOVA					
Reading					
Between groups	11515	1	11515	2.8	0.10
Within groups	217732	53	2400		
Subgroup Comparisons	**Mean**	**sd**	**n**		
1. Low DA not placed	534.29	68.99	17		
2. High DA	565.61	61.99	38		
ANOVA			Mean		
Math	**SS**	**df**	**Square**	**F**	**p**
Between groups	2892	1	2892	1.56	0.21
Within groups	97914	53	931		
Subgroup Comparisons	**Mean**	**sd**	**n**		
1. Low DA not placed	557.17	52.33	17		
2. High DA	572.86	38.24	38		
ANOVA			Mean		
Grade 1 Achievement	**SS**	**df**	**Square**	**F**	**p**
Between groups	1289	1	1289	3.37	0.07
Within groups	20267	53	382		
Subgroups	**Mean**	**sd**	**n**		
1. Low DA not placed	59.47	23.09	17		
2. High DA	69.94	17.8	38		

Table 7. Treatment Validity: ANOVAs showing Subgroup Performance in First Grade - Treatment Sample

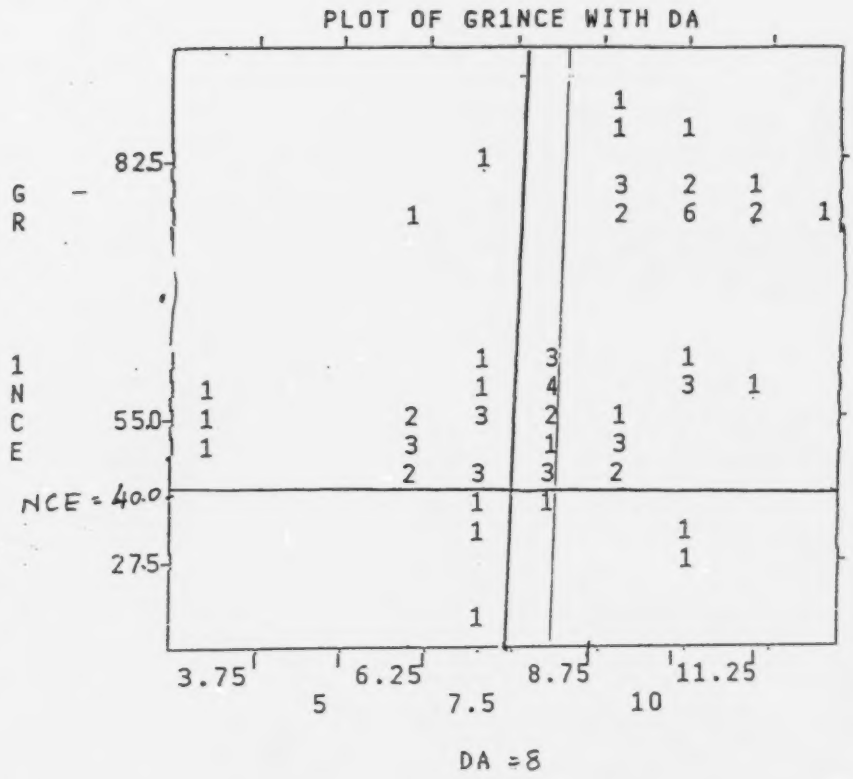
	Sum of Squares	df	Mean Square	F	p
ANOVA: Reading					
Between groups	29066	2	14533	6.05	0.0039
Within groups	156016	67	2400		
Subgroup Comparisons	**Mean**	**sd**	**n**		
1. Low DA placed	510.61	43.72	21		
2. Low DA not placed	535.75	43.50	16		
3. High DA	558.68	54.58*	33		
ANOVA: Math	**SS**	**df**	**MS**	**F**	**p**
Between groups	16386	2	8193	8.79	0.0004
Within groups	60554	67	931		
Subgroup Comparisons	**Mean**	**sd**	**n**		
1. Low DA placed	533.71	28.41	21		
2. Low DA not placed	535.62	16.72	16		
3. High DA	565.67	36.61**	33		
ANOVA: Grade 1 Achievement	**SS**	**df**	**MS**	**F**	**p**
Between groups	333445	2	166722	8.82	0.0004
Within groups	1228767	67	18904		
Subgroup Comparisons	**Mean**	**sd**	**n**		
1. Low DA placed	51.04	13.86	21		
2. Low DA not placed	54.52	9.33	16		
3. High DA	66.36	15.42**	33		

* Significantly different from subgroup 1 at $p \leq .05$
** Significantly different from subgroup 1 & 2 at $p \leq .05$

In the treatment sample (Table 7), the overall ANOVAs yielded significant F ratios, possibly because of the wide gap in performance between the students with initially High DAs versus those in the Low DA category. Students with initially High DAs were significantly and consistently better performers than those with Low DAs. Achievement means of the two subgroups with initially Low DAs (i.e., students who were in the two-year program versus those who were passed on to a one-year kindergarten) were not statistically different in either reading or math. A comparison on the overall first grade measure shows a difference of 3.5 NCE points. Students who started with the lowest DAs were found to remain the lowest achievers in first grade, despite the two-year programming in kindergarten.

Classification Accuracy

Bivariate scatterplots of overall first grade achievement and DA measured at the time of kindergarten entrance are given in Figures 2 and 3, and pertain to data from the treatment and control sample, respectively. In the treatment sample, 24 out of 70 students were initially selected by the GSRT for two-year tracking-- a selection ratio of 34%. If a base rate of 30 % is used along with the empirically-derived predictive validity coefficient of about .40 in the treatment sample, the success ratio in classification is found to be 47% according to the Taylor-Russell tables. This indicates that the test enables 17% more students to be accurately classified over and above the 30% who would have been identified if the test were not used. The treatment sample data suggest that a large number of students initially selected by the test as delayed (88%), eventually perform above the cutoff score (NCE=40) on the criterion, i.e., they are false negatives. In this sample, it may be argued that their improved performance was caused by the two-year intervention given to them prior to first grade. Examination of the scatterplot in the control sample, however, suggests that false negative errors are large on the GSRT even when the recommended intervention is withheld for students in the Low DA category.

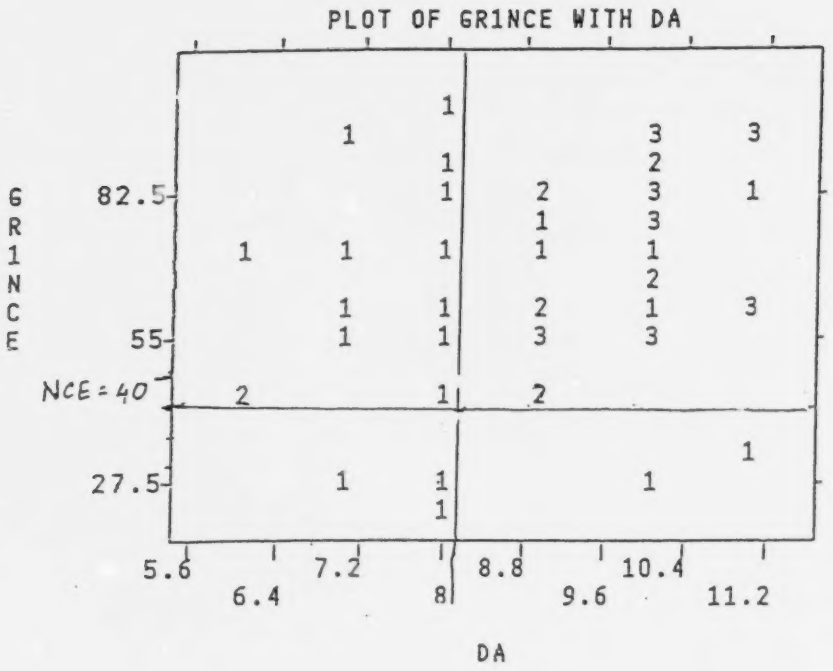


5-5 1/2 years

n = 70
Number of students with low DA's = 24
Selection ratio = 24/70 = 34%
Base Rate = 30%
Predictive Validity coefficient = .39
Success Ratio by Taylor-Russell table = 47%
False Negative Classifications = 21/24 = 88%

Note:The two subgroups with Low DAs, placed vs. not placed, are mixed in the two left qudrants of this plot.

Figure 2. Classification Accuracy: Bivariate Scatterplot of DA with First Grade Achievement in Treatment Sample.



5-5 1/2 year

n = 55
Number of students with low DA's = 17
Selection ratio = 17/55 = 31%
Base Rate = 30%
Predictive Validity coefficient = .29
Success Ratio by Taylor-Russell table = 43%
False Negative Classifications = 14/17 = 82%

Figure 3. Classification Accuracy: Bivariate Scatterplot of DA with First Grade Achievement in Control Sample.

In Figure 3, it is seen that the selection ratio in the control data is 31% (17/55). Using a base rate of 30% and the predictive validity coefficient of about .30 from the control sample, a success ratio in classification of 43% is obtained. This provides a 13% improvement in classification over the base rate. The proportion of false negative errors in the control data is 82% (14/17).

Discussion

This study attempted to examine the predictive properties of the GSRT using the unified approach recommended in the *Standards* (1985). Longitudinal data gathered from different program contexts were analyzed using three approaches. Questions regarding the correlational properties of the test, the efficacy of treatment related to the test, and its classificational properties were investigated. In this final analysis, an attempt is made to integrate the results and make recommendations for the application of the test with reference to its content and psychometric properties.

The GSRT was found to be a positive but modest predictor of achievement measured in first grade. Zero-order correlations were in the range of .24 -.39, indicating an overlap in variance of 6 -15%. Correlations in the treatment program in this study were found to be higher than those in the control program, an outcome that may have resulted from the different ranges of DAs in the two samples, as well as the different sample sizes. Differential intervention also affected the overall variability in achievement within the two groups, yielding correlations of different magnitude. Other studies that have reported correlations of DA with standardized achievement that are of similar magnitude are those by Popovics (1982), May (1986) and Graue and Shepard (1988).

Incremental validity analyses yielded rather similar results with data from both treatment and control samples. DA was found to add substantially and significantly to the prediction of achievement when used as an added predictor with CA. This was true in all the analyses which used math and reading scores as separate dependent variables, as well as when the combined achievement measure served as the dependent variable. There was a small positive correlation of chronological age with achievement. Percent of variance in achievement predicted with both variables in the prediction equations was often larger than when CA was eliminated. However, the R squared values with both CA and DA in the equations were still modest in magnitude (.08 - .16). The observed overlap of DA with achievement obviously stems from the content of the GSRT, which contains school-related tasks such writing and oral language competency. Wood et al (1982) compared CA with DA as predictors of future school success and reached similar conclusions. The outcome that DA is a better predictor of achievement than CA supports the same claim made by the authors of the test.

The analyses on treatment validity within the two samples in this study suggested that students with initially high DAs do better on subsequent achievement measures than students with initially low DAs. The observed

positive correlations of DA with achievement lends support to this inference. Given the actions that follow test use, the important question to consider is whether the observed positive correlations are adequate evidence to justify the application of the GSRT in conjunction with the recommended treatment. In other words, are students with Low DAs benefitting from the treatment of extended kindergarten programs? The data from both samples in this study indicated that the relative rank ordering of students with low and high DAs was found to remain unchanged on first grade achievement, irrespective of the type of programming given to delayed students in the interim period. In other words, even when the recommended two-year treatment was given to students with initially the lowest DAs, their performance stayed the lowest in first grade in comparison with other students who had a one-year kindergarten. It may be argued that the students with the lowest DAs probably did better than they would have done without the recommended treatment of an extra growth year. The data, however, suggested that the extra-year treatment was not effective in enabling students in the lowest DA category to catch up with their first grade peers, even when they were a year older than the latter group. It is possible that the GSRT screens students with special needs who may benefit from specialized treatment that goes beyond the provision of an extra year. The findings on treatment validity concur with those reported by May and Welch (1984).

Examination of classification properties of the GSRT indicated that, with the given base rate and selection ratio, and observed correlation with achievement of around .30, the test would allow successful classification of 43% of the students--an improvement of only 13% over the percent identified on the criterion alone (i.e., without use of the test). False negative errors were rather high. Although the proportion of observed false negative errors is a function of the cutoffs employed on the predictor and criterion scales. The outcomes may also point to the possibility that when students starting with the lowest DAs are given appropriate learning opportunities in the classroom (rather than simply the extra year), they are no longer in the at-risk category by the end of first grade. The appropriate treatment for such students, therefore, may not be the additional time in kindergarten as it may be the educational experiences that promote growth and development.

In summary, the GSRT was found to be a modest predictor of subsequent school achievement, and a better predictor than chronological age. However, the predictive validity coefficients alone were not sufficient evidence to support the application of the test in placing and extra-year tracking of students with low DAs. Furthermore, the classificational properties of the GSRT were found to be rather weak, yielding a large proportion of false negative errors in identifying at-

risk students on the low end of the scale.

Admittedly, the sample sizes used in this study may be a cause for concern. It must be noted in this regard that several recently completed studies already cited showed very similar results. The ethnic composition of the samples with predominantly white students may limit the generalizability of the outcomes to districts with a similar ethnic make-up. In the school district where the study was conducted, however, which has a predominantly white rural population, the make-up of the samples compared well with the ethnic composition of the elementary level student population.

Recommendations

The evidence from this study failed to support the use of the GSRT as a placement tool for extra-year programs. As the content of the GSRT indicates, it is an early childhood screening test. The test may therefore be used as an initial screener only, as recommended by Miesels (1983), rather than as a placement or diagnostic tool. As a screening test, its use should be followed by formal diagnostic assessments prior to making placement decisions. Recommendations for extra-year treatments should be avoided for students with low performance on the GSRT. Teachers may explore the utility of the profile from the screening in instructional planning for entering kindergartners.

The distinctions between screening, readiness, and diagnostic tests should be made clear to users by the test developers to avoid future applications of the GSRT (or subtests of it) in placement and diagnosis.

REFERENCES

AERA, APA and NCME. (1985). Standards for educational and psychological testing. Washington DC: Author.

Ames, L.B., Gillespie, C., Haines, J., & Ilg, F.L. (1979). The Gesell Institute's child from one to six: evaluating behavior of the preschool child. New York, NY: Harper and Row.

Ames, L.B., & Ilg, F. L. (1964). Gesell behavior tests as predictive of later grade placement. Perceptual and Motor Skills, 19(719-722).

Angoff, W. (1988). Validity: An evolving concept. In Test Validity by H. Wainer and H.I. Braun (Editors). Hillsdale, NJ: Lawrence Erlbaum Associates Inc., Publishers.

Banerji, M. (1990). Validation and Psychometric Analysis of the Gesell School Readiness Screening Test under Local Conditions. Unpublished Dissertation. University of South Florida, College of Education.

Bradley, R.H. (1985). Review of Gesell School Readiness Test. In the Ninth Mental Measurements Yearbook, Volume 1 by James V. Mitchell (Ed.). Lincoln, Nebraska: The Buros Institute of Mental Measurements.

Bredekamp, S. and Shepard, L. A. (1989). How best to protect children from inappropriate expectations, practices, and policies. Young Children, 44, (14-24).

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York, NY: Holt, Reinhart and Winston.

Cronbach, L. J., & Meehl, P. (1955). Construct validity in psychological tests. Psychological Bulletin, 52(281-302).

District School Board of Pasco County. (1987). Kindergarten Curriculum Guides. Land O' Lakes, FL: Author.

Florida Department of Education, Division of Public Schools. (1989). 1989 Survey of Alternate Placement Programs in Florida - Kindergarten through Grade 3. Tallahassee, FL: Author

Gesell Institute of Human Development. (1980). The Gift of Time...the developmental point of view. New Haven, CT: Author.

Graue, Elizabeth M. and Shepard, Lorrie A. (1988). Predictive Validity of the Gesell School Readiness Test. ERIC Database.

Guion, R. M. (1976). Recruiting, Selection, and Job Placement. In M. D. Dunnette (Ed.) Handbook of Industrial and Organizational Psychology. Chicago: Rand McNally.

Haines, J., Ames, L.B., & Gillespie, C. (1980). The Gesell Preschool Test Manual. Lumberville, PA: Modern Learning Press.

Ilg, F. L., & Ames, L.B. (1972). School Readiness. New York: Harper and Row.

Karweit, Nancy L. (1988). Effective kindergarten programs and practices for students at risk. In Effective Programs for Students At Risk by R. E. Slavin, N. L. Karweit and Nancy A. Madden (Editors). Boston, MA: Allyn and Bacon

Kaufman, A. S. (1971). Piaget and Gesell: A psychometric analysis of tests built from their tasks. Child Development, 42 (1341-1360).

Kaufman, A. S. & Kaufman, N. L. (1972). Tests build from Piaget's and Gesell's tasks as predictors of first grade achievement. Child Development, 43 (521-535).

Kaufman, N. L. (1985). Review of the Gesell Preschool Test. In The Ninth Mental Measurements Yearbook, Volume 1 by James V. Mitchell (Ed.). Lincoln, Nebraska: The Buros Institute of Mental Measurements.

May, D.C. (1986). Relationships between the Gesell School Readiness Test and standardized achievement and intelligence measures. Educational and Psychological Measurement, 46(1051-1059).

May, D. C., & Welch, E. (1984). The effects of developmental placement and early retention on children's later scores on standardized tests. Psychology in the Schools, 21 (381-385).

Messick, S. (1989). Validity. In Educational Measurement by R. L. Linn (Editor). New York: American Council on Education.

Meisels, S. J. (1983). Developmental screening in early childhood: A guide. Washington DC: NAEYC.

Meisels, S. J. (1987). Uses and Abuses of Developmental Screening and School Readiness Testing. Young Children, 42(2).

Meisels, S. J. (1988). Personal Communication.

Naglieri, J. A. (1985). Review of Gesell Preschool Test. In the Ninth Mental Measurements Yearbook, Vol. 1 by James V. Mitchell, Jr. (Editor). Lincoln, Nebraska: The Buros Institute of Mental Measurements.

Popovics, A. J. (1982). Total and component actuarial scores of the Gesell Copy Forms test as predictors of achievement, intelligence, and creativity measures. The Journal of Psychology, 110, (293-295).

Sechrist, Lee. (1963). Incremental Validity: A Recommendation Educational and Psychological Measurement, 23, (153-158)

Shepard, L. A., & Smith, M. L. (1985). Boulder Valley Kindergarten Study: Retention Practices and Retention Effects. Boulder, Colorado: Laboratory of Educational Research.

Shepard, L. A., & Smith, M. L. (1986). Synthesis of research on school readiness and kindergarten retention. Educational Leadership, 44 (78-86).

Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and Tables. Journal of Applied Psychology, 23, (565-578).

Waters, E. (1985). Review of the Gesell School Readiness Test. The Ninth Mental Measurements Yearbook. Vol. 1 by James V. Mitchell, Jr. (Editor). Lincoln, Nebraska: The Buros Institute of Mental Measurements.

Wood, C., Powell, S., & Knight, R. C. (1984). Predicting school readiness: the validity of developmental age. Journal of Learning Disabilities, 17(8-11).